DOCUMENT RESUME

ED 414 336 TM 027 869

AUTHOR Crehan, Kevin D.
TITLE A Discussion of Analytic Scoring for Writing Performance Assessments.
PUB DATE 1997-10-00
NOTE 10p.; Paper presented at the Annual Meeting of the Arizona Educational Research Association (Phoenix, AZ, October 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Methods; Feedback; *Generalizability Theory; *Interrater Reliability; *Performance Based Assessment; Scoring; *Test Reliability; Writing (Composition); *Writing Tests
IDENTIFIERS *Analytic Scoring; Scoring Rubrics

ABSTRACT

Writing fits well within the realm of outcomes suitable for observation by performance assessments. Studies of the reliability of performance assessments have suggested that interrater reliability can be consistently high. Scoring consistency, however, is only one aspect of quality in decisions based on assessment results. Another is generalizability. Research suggests that if the number of ratings per task could be increased, it may yield an increase in "task" generalizability without a dramatic increase in the actual number of tasks. Multitrait analytic scoring strategies for writing performance assessments may increase "task" generalizability over a single holistic score. Research undertaken by G. Roid (1994) supports the potential usefulness of analytic scores as effective sources for feedback to students and as bases for meaningful discussion on the writing process. Work at the Center for the Study of Evaluation at the University of California, Los Angeles, has expanded on the development of methodology and uses of analytic scoring. Work on narrative-writing-specific scoring rubrics has shown promising evidence of reliability and validity. Training in and use of these rubrics has also increased participating teachers' understanding of the quality components of writing. (Contains 3 figures and 15 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 414 336

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A Discussion of Analytic Scoring for Writing Performance Assessments

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Kevin Crehan
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Kevin D. Crehan

University of Nevada, Las Vegas

The most prevalent response to the call for assessment reform has been to increase the use of more authentic assessments, e.g., performance assessments. Advocates of performance assessments suggest that this from of appraisal can serve to measure important and complex learning outcomes and provide information useful to guide improvement in instruction (Resnick & Resnick, 1989). Perhaps the most complex form of student achievement which we attempt to assess involves composition. Therefore, the task of writing fits well within the realm of outcomes suitable for observation by performance assessments.

Among the problems associated with using performance assessments to measure important learning outcomes are objectivity of ratings and generalizability (reliability) of scores across raters and tasks. A review by Linn (1993) summarized evidence of acceptable generalizability across raters given well-defined scoring rubrics, intensive

Paper presented at the annual meeting of the Arizona Educational Research Association, Phoenix, AZ, October, 1997

TMO27869

BEST COPY AVAILABLE

rater training, and monitoring during rating. Additionally, the California Assessment Program has established an inter-rater reliability of .90 for their writing assessment by using procedures which include providing sample anchor papers for each rater and recirculating previously scored papers to check on stability (U.S. Congress, Office of Technology Assessment, 1992). Shavelson, Baxter, and Pine (1992) observed the reliability and validity of performance assessments in the 5th and 6th grade science curriculum. They asked the question: How large a sample of observers is needed to produce reliable measurement? Their results found inter-rater reliability to be consistently high in evaluating student performance on complex tasks, high enough to conclude that a single rater provides a reliable score.

While these observations offer promise for the utility of performance assessments, scoring consistency is only one aspect of quality in decision situations based on assessment results. Linn and Burton (1994) suggest that for pass-fail decisions involving individual students, acceptable generalizability across tasks is attained only when a large number of tasks are used, perhaps as many as ten. If the content aera is being assessed in writing, such a large number of writing tasks on an occasion might require an unreasonable expenditure of instructional time devoted to assessment to say nothing of the administration and scoring costs. However, if the number of ratings per task could be increased, it may yield an increase in "task" generalizability without a dramatic increase in the actual number of tasks. Multitrait analytic scoring strategies for writing performance assessments may increase "task" generalizability over a single holistic score.

Much of the research on the psychometric characteristics of writing performance

assessments uses single score "holistic" ratings. In writing assessment this single holistic score designed to estimate the wholeness in quality of the writing product. There is agreement (e.g., Huot, 1990) that writing is a multifaceted performance and as such involves attainment on a number of mental traits, e.g., vocabulary, language mechanics (see Figure 1), on which individual differences exist. Additionally, there are different types of writing, e.g., narrative, expository (see Figure 2). Given that writing performance involves a number of traits on which individuals differ, analytic scoring of writing products is recommended by some researchers (see Figure 3) (Roid, 1994; Huot, 1990; Marsh & Ireland, 1987; Novak, Herman, & Gearhart, 1996).

Roid (1994) used cluster analyses to explore the empirical validity of the analytic traits presented in Figure 1. Results of these analyses demonstrated that, while forty percent of the responses had flat trait patterns (either all high or low), a number of distinct patterns among the six traits were evidenced. For example, thirteen percent of the patterns were very close to average on five of the traits but either high or low on conventions. Ten percent of the patterns showed high or low voice, with other scores near average. An additional thirteen percent were either high or low on ideas, organization, and voice but close to average on word choice, sentence fluency, and conventions. This suggests evidence of a creative or stylistic component among the six traits. This evidence supports the potential usefulness of analytic scores as effective sources for feedback to students and as bases for meaningful discussion on the writing process.

Work at the Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, at UCLA (e.g., Wolf & Gearhart, 1993a;

1993b) has expanded on the development of methodology and uses of analytic scoring. Work on narrative-writing-specific scoring rubrics has shown promising evidence of reliability and validity (Gearhart, Herman, Novak, Wolf, & Abedi, 1994; Gearhart, Herman, & Novak, 1996). Additionally, training and use of these rubrics has benefited instruction by increasing participant teachers' understanding of the quality components of writing (Gearhart & Wolf, 1994; Gearhart et al., 1994, Wolf & Gearhart, 1995).

References

Gearhart, M., & Wolf, S.A. (1994). Engaging teachers in assessment of their students' writing: The role of subject matter knowledge. Assessing Writing, 1, 67-90.

Gearhart, M., Herman, J. L., Novak, J. R., Wolf, S. A., & Abedi, J. (1994). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric (CSE Tech. Rep. 389). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Herman, J. L., & Novak, J. R. (1996). Issues in portfolio assessment: The scorability of narrative collections. (CSE Tech. Rep.) Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Wolf, S. A., Burkey, B., & Whittaker, A. K. (1994). Engaging teachers in assessment of their students' narrative writing: Impact on teachers' knowledge and practice (CSE Tech. Rep. 377). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. Review of Educational Research, 60, 237-263.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. Educational Evaluation and Policy Analysis, 15, 1-16.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. Educational Measurement: Issues and Practice, 13, 5-8, 15.

Resnick, L. B., & Resnick, D. P. (1989). Assessing the thinking curriculum: New

tools for educational reform. In B.R. Gifford & M.C. O'Conner (Eds.), Future Assessments: Changing views of aptitude, achievement, and instruction (pp.37-75). Boston, MA: Kluwer.

Roid, G. H. (1994). Patterns of writing skills derived form cluster analysis of direct-writing assessments. Applied Measurement in Education, 7(2), 159-170.

Shavelson, R. J., Baxter, G. P., & Pine J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher,21,22-27.

Spandel, V., & Stiggins, R. J. (1990). Creating writers: Linking assessment and writing instruction. White Plains, NY: Longman. instructional decisions.

U.S. Congress, Office of Technology Assessment. (1992) Testing in American Schools: Asking the Right Questions (OTA-SET-519), Washington, D.C.: U.S. Government Printing Office.

Wolf, S. A., & Gearhart, M. (1993a). Writing What You Read: Assessment as a learning event (CSE Technical Report 358). Los Angeles: University of California. Center for Research on Evaluation, Standards, and Student Testing.

Wolf, S. A., & Gearhart, M. (1993b). Writing What You Read: A guidebook for the assessment of children's narratives (CSE Resource Paper No. 10). Los Angeles: University of California. Center for Research on Evaluation, Standards, and Student Testing.

Wolf, S. A., & Gearhart, M. (1995). Engaging teachers in assessment of their students' narrative writing: Patterns of impact and implications for professional development. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Figure 1

Definitions of Analytic Traits (Spandel & Stiggins, 1994)

Ideas	**The heart of the message, the content of piece, the main theme, together with all the details that enrich and develop that theme. Ideas are strong when the message is clear and enlivened with interesting and important details.**
Organ-ization	**The internal structure of a piece of writing, the thread of central meaning, the pattern that holds everything together. Organization is strong when the piece begins meaningfully, proceeds logically, and creates a sense of anticipation that is ultimately systematically fulfilled.**
Voice	**The writer coming through the words, his or her wit and feeling, the sense that a real person is speaking to us and cares about the message. Good writers impart a personal tone and flavor to the piece that is unmistakably his or her's alone.**
Word Choice	**The use of rich, colorful, precise language that communicates not just in a functional way but in a way that moves and enlightens the reader. Strong word choice may depend more on the skill of using words precisely than on an exceptional vocabulary.**
Sentence Fluency	**The rhythm and flow of the language, the sound of word patterns, the way in which the writing plays to the ear - not just to the eye. With good fluency, sentences vary in length and style, and they are so well-crafted that reading aloud is a pleasure.**
Conven-tions	**The mechanical correctness of the piece - spelling, grammar, usage, paragraphing, capitalization , and punctuation. Writing that is strong in convention has been well proofread and edited.**

Figure 2

Modes of Writing (Roid, 1994)

Descriptive	**Describes an object, place, or person, enabling the reader to visualize what is being described and to feel that he or she is very much part of the writer's experience. Writer's purpose is to create a strong and vivid image of impression in the reader's mind.**
Persuasive	**Attempts to convince the reader that a point of view is valid or persuade the reader to take a specific action. Writer's purpose is to persuade the reader.**
Expository	**Gives information, explains something, clarifies a process, or defines a concept, Writer's purpose is to inform, clarify, explain, define, or instruct.**
Narrative	**Recounts a personal experience or tells a story based on a real event. Writer's purpose is to recount an experience or tell a story in a concise and focused way to create some central theme or impression in the reader's mind.**
Imaginative	**Tells a story based on the writer's imagination. The story is basically fictional, but the writer may use his or her experience and knowledge of people or situations to bring a special flair or flavor to the writing. Writer's purpose is to entertain the reader or write for the author's own pleasure.**

Figure 3

Advantages and Limitations of Multifaceted Analytic Scoring (Gearhart & Wolf, 1994; Gearhart, Wolf, Burkey, & Whittaker, 1994; Spandel & Stiggins, 1990; Wolf & Gearhart, 1995)

Advantages:

1. **Developing the analytic scoring rules forces judgements on what is valued in writing and the product provides an operational definition for the quality characteristics of writing.**

2. **Allows more systematic and detailed feedback to students on the strengths and weaknesses of their writing.**

3. **Provides more diagnostic information that teachers may use to guide their instruction and student practice.**

4. **Benefits the teachers who are trained in the rating method and subsequently perform the ratings. These teachers can use what they learn to improve their writing instruction and feedback to students.**

5. **Ratings on multiple facets of the domain of writing skills allows improved generalizability over a single holistic score.**

Limitations:

1. **Analytic scoring can be very expensive and time consuming if not well managed.**

2. **The analytic rating task is not for everybody. The rating task is initially difficult and beginning raters may experience frustration.**